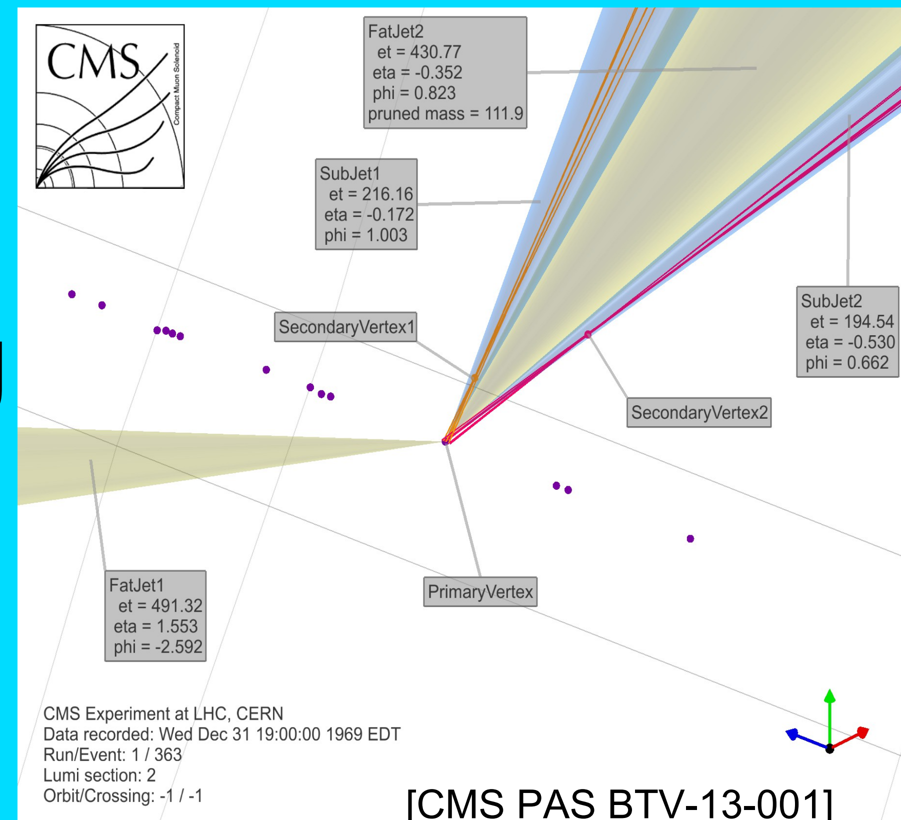
Heavy Flavour Tagging at CMS



[CMS PAS BTV-13-001]

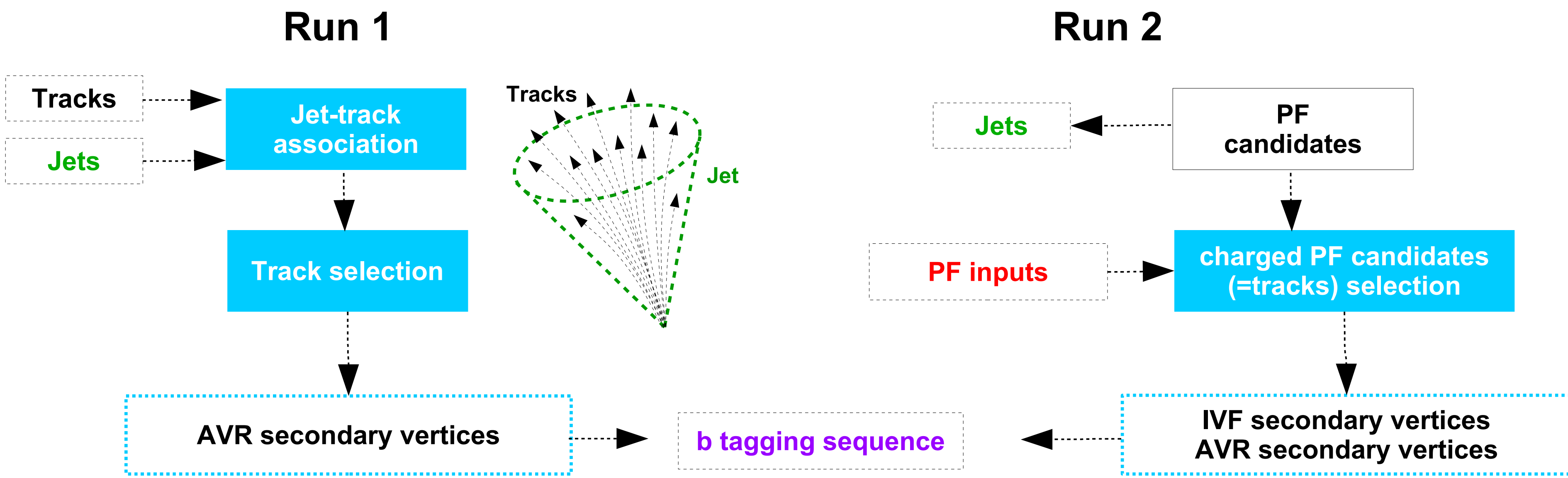
Ivan Marchesini, HF-LHC2016, 21 Apr 2016

on behalf of the CMS Collaboration

b tagging at CMS

results from **CMS PAS BTV-15-001**
http://cms-results.web.cern.ch/cms-results/public-results/preliminary-results/BTV-15-001/index.html

Run 2 b tagging



- Restructured framework in Run 2:
  - b tagging workflow directly interfaced to **Particle Flow (PF)** reconstruction: better exploit PF information, e.g. **K0** and **nuclear interactions** reconstruction, fake rejection
- New algorithm to reconstruct secondary vertices (SVs):
  - Run 1: **adaptive vertex reconstruction (AVR)**, starts SV fit from jet tracks
  - Run 2: default algorithms is the **inclusive vertex finder (IVF)**, starts from **all tracks in the event**, no prior jet-track association. Essential for **double b jets**
    seeds for SV fit are displaced tracks with IP>50 μm and IP significance > 1.2

- AK5 → **AK4** jets

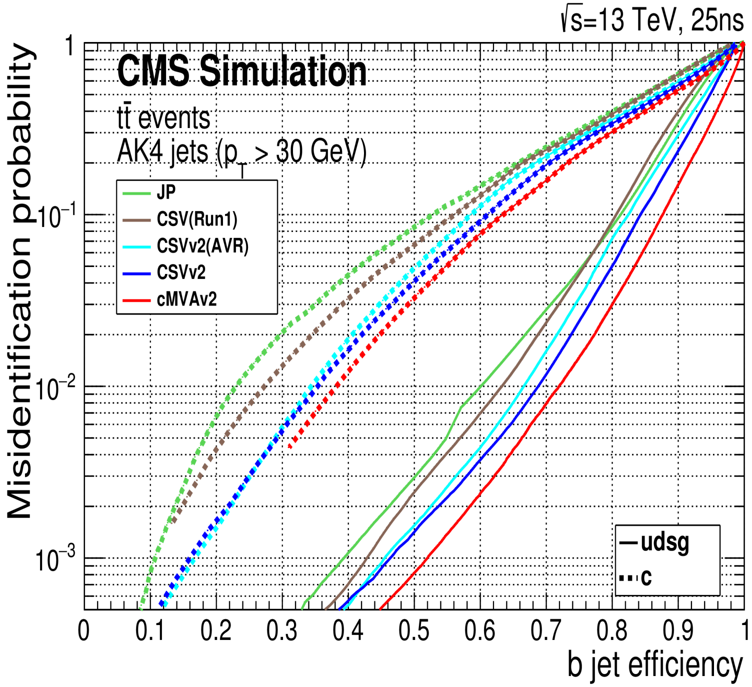
Run 2 taggers

- **Combined Secondary Vertex CSV**, flagship tagger for Run 1, exploits:
  - displaced tracks
  - AVR secondary vertices

- CSV algorithm significantly improved→**CSVv2**:
  - **neural network** instead of a Likelihood Ratio
  - additional variables, improved track selection
  - use of **IVF** secondary vertices

- **Jet Probability (JP)** algorithm:
  - mostly used for performance measurements
  - based on track displacement
  - calibrated separately in data and MC using tracks with negative IP

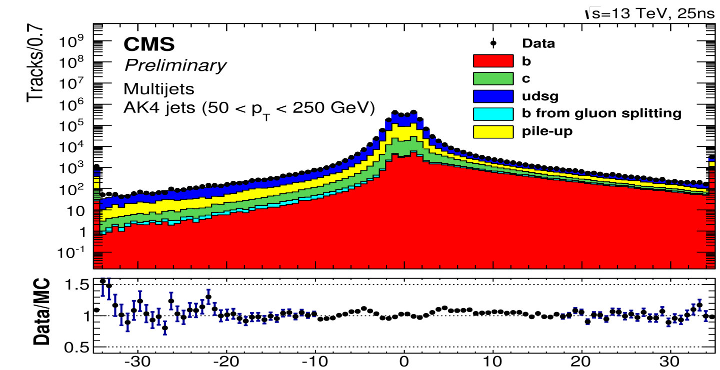
CSV: Run 1 → Run 2



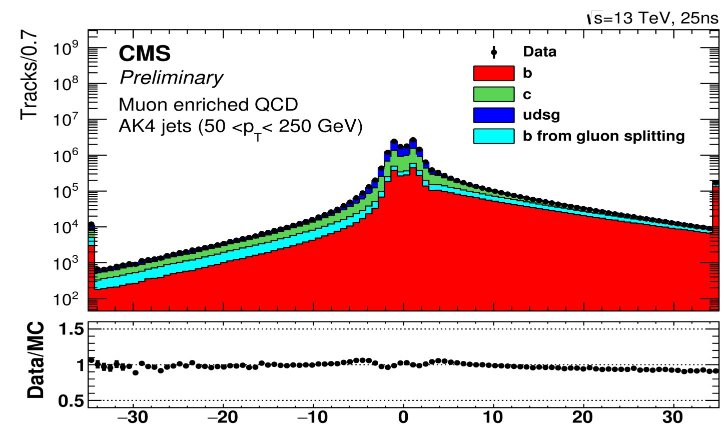
- **cMVAv2** algorithm
  - new algorithm developed in Run 2
  - it combines in a boosted decision tree (BDT) the discriminators from other algorithms:
    - **JP** taggers, **CSVv2(IVF)** and **CSVv2(AVR)**
    - Soft Muon (**SM**) and Soft Electron (**SE**) taggers: soft lepton kinematic observables

Commissioning: three topologies

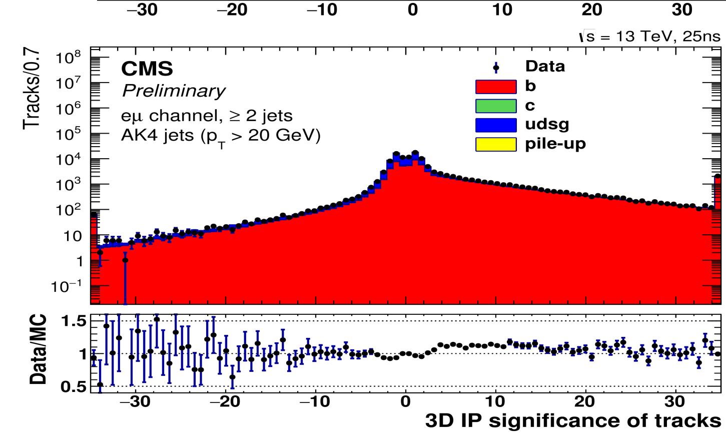
- **Inclusive QCD** multijet events, enriched in **light flavor**:
  - ➔ mistag scale factors
  - data collected by inclusive jet triggers



- **Muon-enriched QCD** multijet topology, enriched in **b flavor,** jets containing **soft muon** p_T>5 GeV:
  - ➔ b tagging scale factors
  - data collected by dedicated calibration triggers, jets with muon



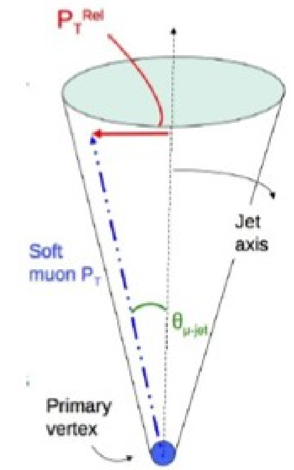
- **Dilepton ttbar**: events with **two jets** (p_T>20 GeV) and a pair of opposite charge **isolated leptons** (p_T>20 GeV), enriched in **b flavor**:
  - ➔ b tagging scale factors
  - ➔ discriminator re-weighting
  - data collected by dilepton triggers



Performance measurements: overview

- Purpose: correct for data/MC discrepancies in the b tagging performance
  - **scale factors** to correct for specific cuts on the discriminators (**working points**)
  - correction factors for **reshaping the whole discriminator** distribution, for analyses exploiting shape (e.g. in MVA)

- Measurement of the **b tagging efficiency**, based on samples enriched in b jets:
  - jets with a **soft muon** coming from a semileptonic decay of a B hadron
    - **PtRel** method
    - **Lifetime Tagger** method
    - **System8** method
  - dilepton **ttbar** sample
    - **Tag Counting** method

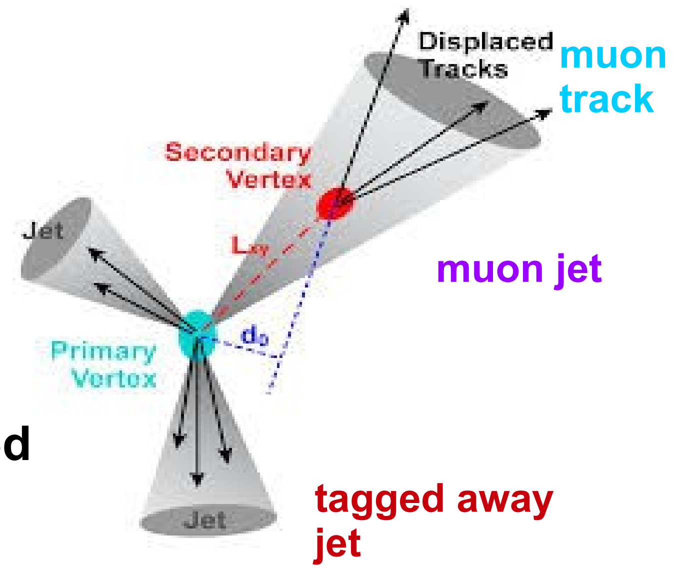


- **Discriminator reweighting**:
  - evaluated both for b jets and light jets
  - both b-enriched (ttbar dilepton) and light-enriched (Z→leptons) samples exploited

- Measurement of the **misidentification probability** for light jets:
  - performed on **inclusive QCD** sample
  - **negative tag** method

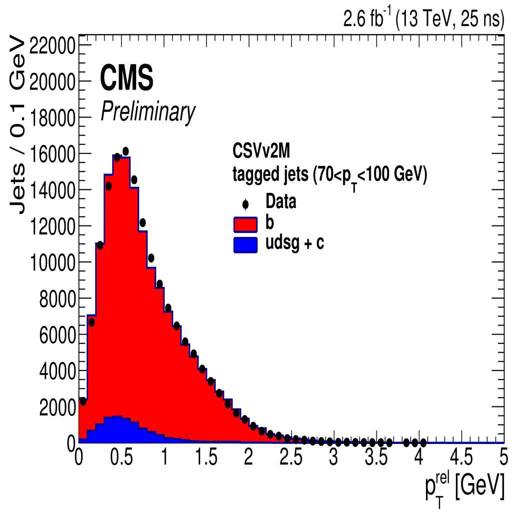
Example: PtRel method

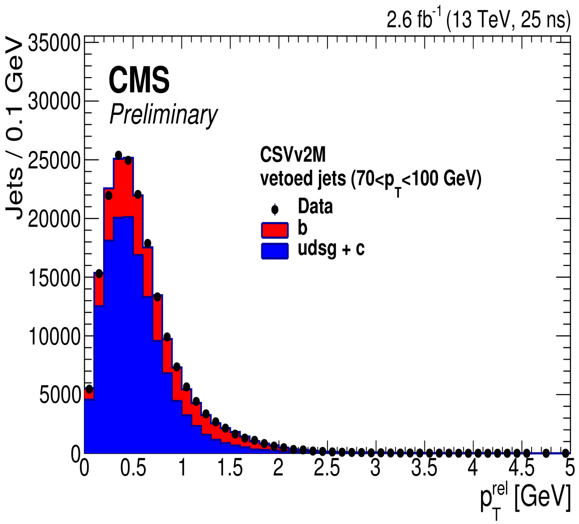
- Require tagged jet (tagged away) in the event to enrich sample in b jets



- Template fit for muon jet:
  - based on p_T^{rel} distribution
  - fit data for fractions of b and c+light jets
  - templates **from simulation**
  - the shapes of the **templates for light jets are corrected** based on the data/MC ratio observed

- Efficiency in data is derived, based on the subsamples of muon jets passing or failing the b tagging requirements

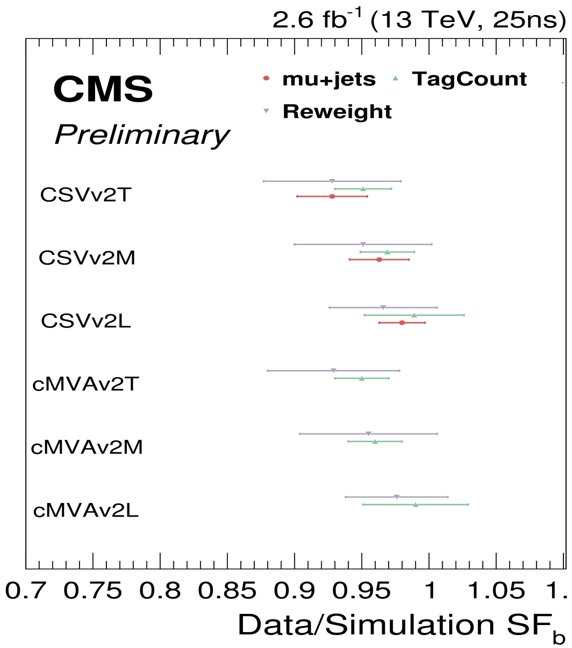




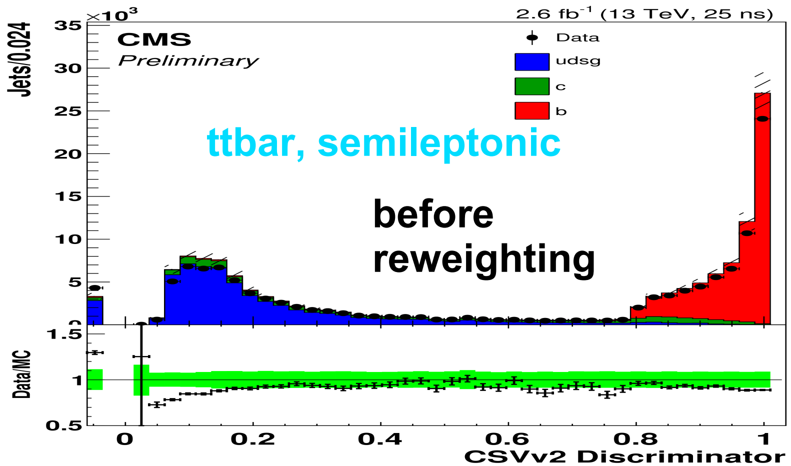
$$\epsilon_b = \frac{N_b^{tagged}}{(N_b^{vetoed} + N_b^{tagged})}$$

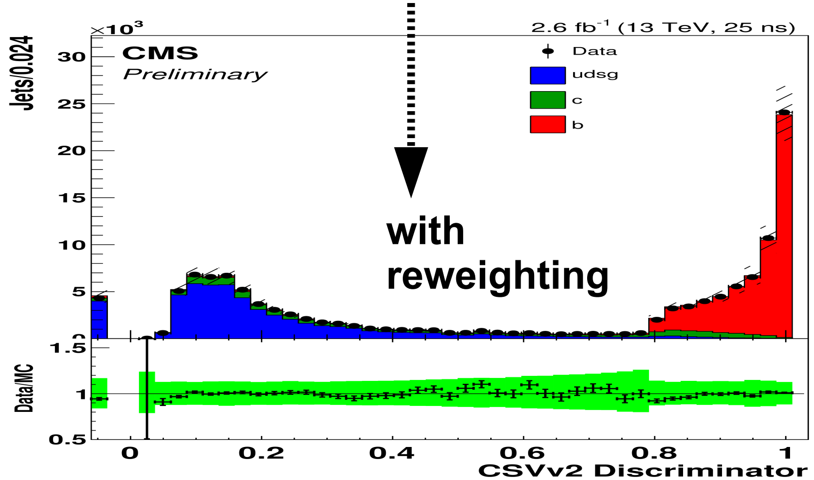
Scale factors

- Consistent results from different techniques and different samples
- Here compared:
  - combined results from **muon-enriched QCD**, averaged over the p_T spectrum of b jets from ttbar
  - **TagCount** method results (ttbar)
  - average scale factors obtained applying the **reweighting** method on ttbar events



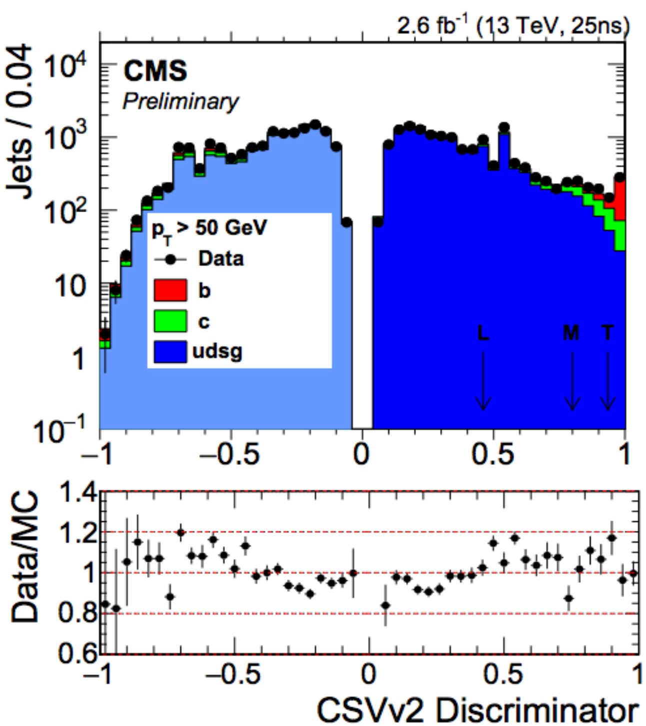
- Discriminator **shape reweighting** closure test
- Good data/MC after SFs are applied

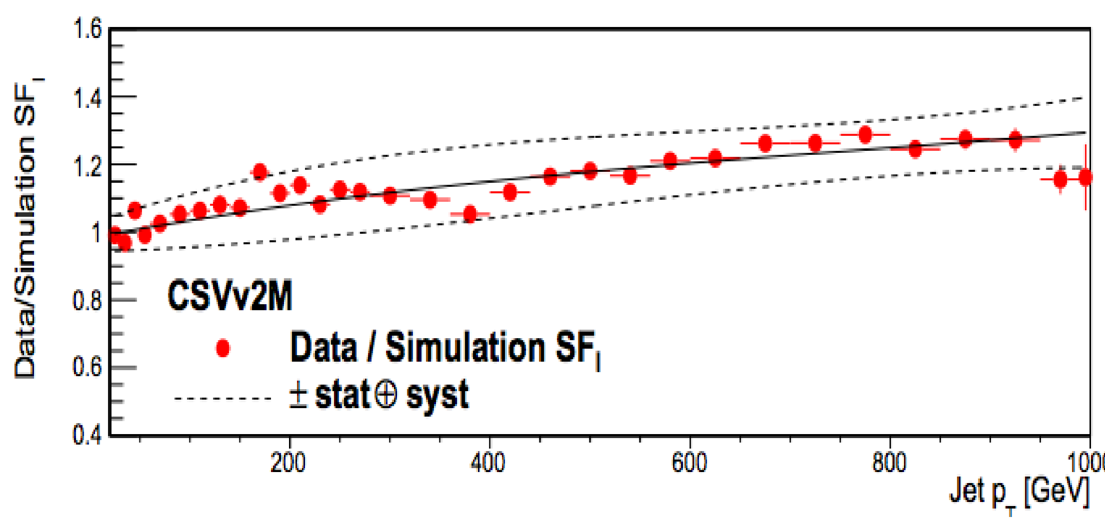




Misidentification probability

- For any tagger, the corresponding **negative tagger** is defined, based on tracks with negative impact parameter
- The negative and positive tag rates are related: $\varepsilon^{mistag} \equiv \varepsilon^{postag}_{udsg} = R_{light}\,\varepsilon^{negtag}_{all}$
- Factor R_{light} :
  - extracted from simulation
  - assigned systematics from negative/positive tag rate asymmetry and heavy flavor contribution



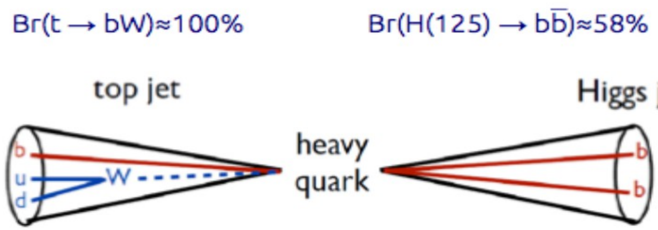


Boosted topologies

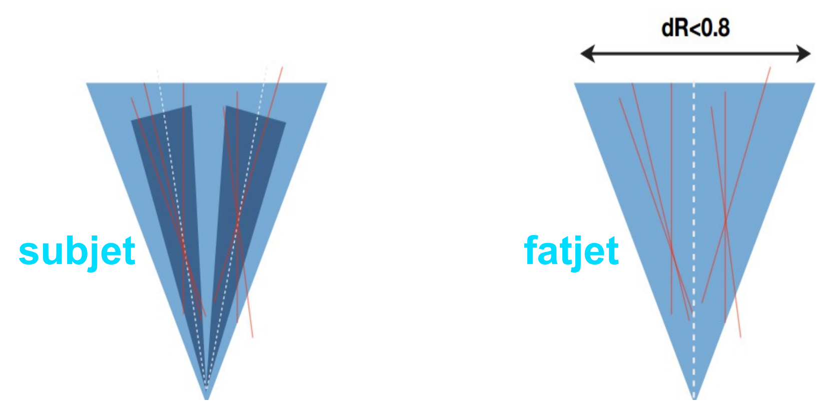
subjet b tagging MC studies form **CMS DP-2014/031**
https://twiki.cern.ch/twiki/bin/view/CMSPublic/BoostedBTaggingPlots2014
boosted double b tagger MC studies from **CMS DP-2015/038**
https://twiki.cern.ch/twiki/bin/view/CMSPublic/BoostedBTaggingPlots2015
commissioning results from **CMS PAS BTV-15-001**
http://cms-results.web.cern.ch/cms-results/public-results/preliminary-results/BTV-15-001/index.html

Boosted b tagging

- **Run 1 success**:
  - employed by **several public analyses**
  - two channels:
    - boosted top
    - boosted Higgs→bb



- **Two approaches**:
  - **subjet b tagging**: Run 1 baseline, b tagging on subjet tracks
  - **fatjet b-tagging**:
    - b tagging uses all jet tracks
    - overall outperfomed by subjet b tagging
    - evolves in Run 2 in dedicated **TMVA-based tagger**, specifically trained for the boosted topology considered:
      - **double-b tagger (Higgs → bb)** tagger (later)
      - boosted top (in preparation)

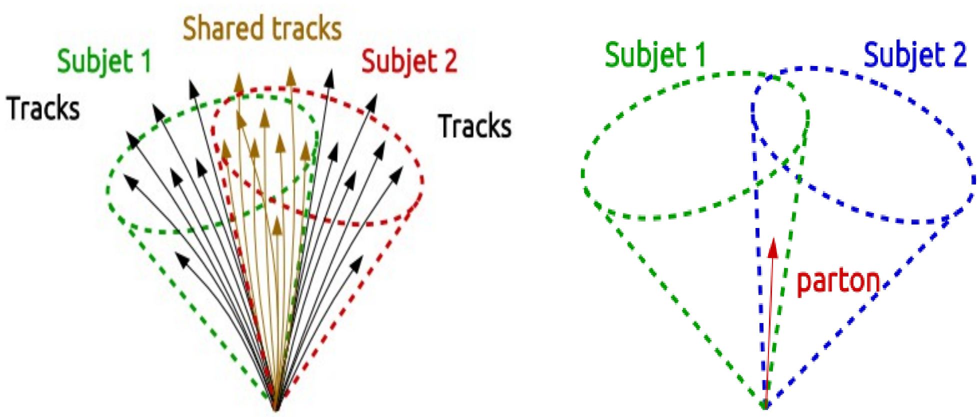


- Orthogonal to **substructure**: can be combined with substructure requirements (n-subjettiness, top-tagging, ...)

Subjet b tagging: Run 2 improvements

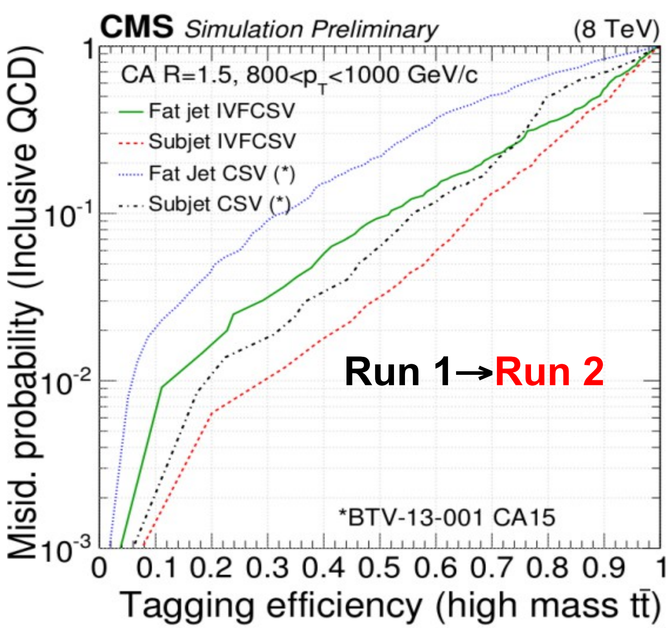
- **Jet-track association**:
  - based on a **fixed-size cone**
  - can lead to double-counting of tracks at high p_T

→ Run 2: use tracks linked to **charged constituents** of particle-flow (sub)jets



- **Jet-flavor assignement**:
  - also based on a fixed-size cone ($\Delta R<0.3$) around gen level parton
  - can lead to subjet flavor ambiguities

→ Run 2: using b and c **hadrons** instead of b and c quarks
→ Run 2: based on **clustering** "ghost" hadrons/partons instead of ΔR matching

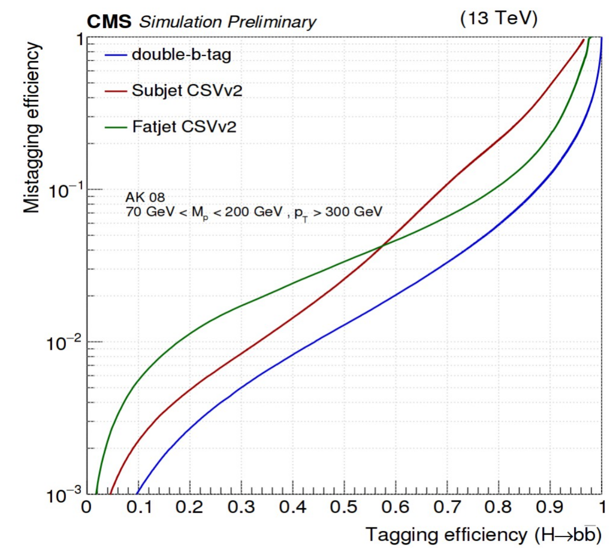


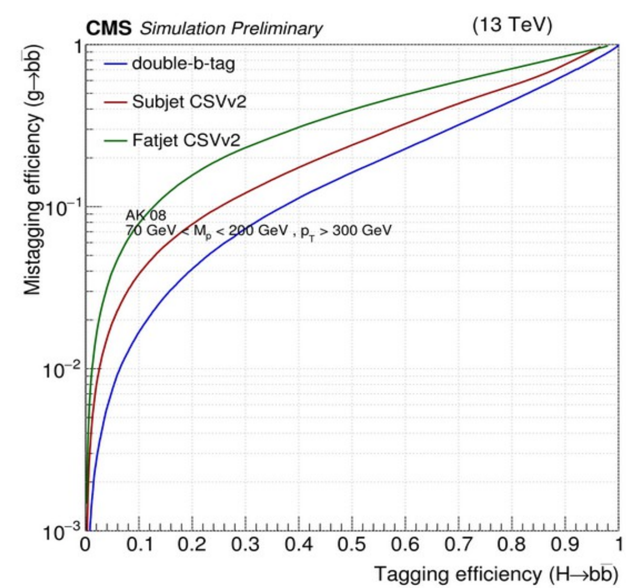
- Other improvements:
  - **IVF** secondary vertices
  - improved **CSVv2** tagger

Boosted TMVA double b tagger

- **New strategy**:
  - multivariate tagger targeting boosted decays to b pairs (e.g. Higgs→bb)
  - stable against p_T, independent from mass of particle
  - two cone sizes:
    - **0.8**: boosted regime
    - **1.5**: low boost regime
- **Training**:
  - BDT training against QCD background
  - information used:
    - track related
    - secondary vertex related
    - minimum CSVv2 subjet score
    - if two SVs found:

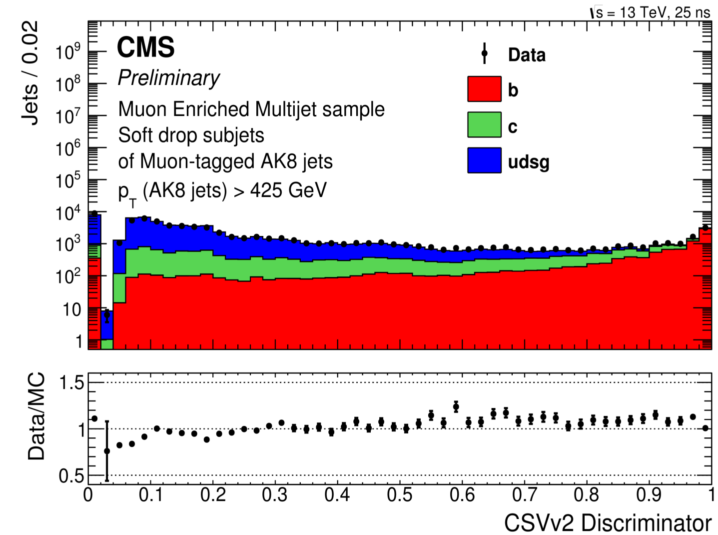
    $Z = dR(SV1,SV2)^{*}\ z$ where $z = p_T 1/mass(SV1 + SV2)$
- Overall outperformes subjet and fatjet b tagging
  - good discrimination also against QCD gluon splitting→bb
- Further improved version exists (released soon)



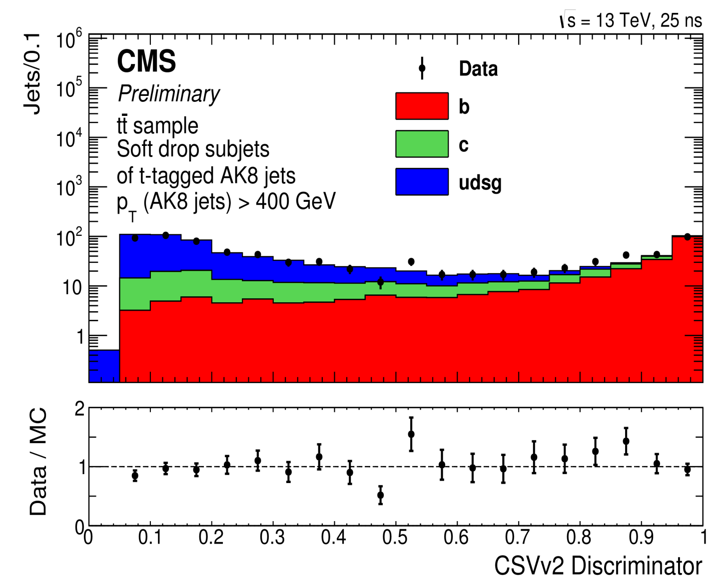


Boosted b tagging commissioning

- Channel 1: boosted **double b tagging**, e.g. H→bb
  - issue: not enough boosted H or Z→bb in data
  - same strategy as in Run1: validation using gluon-splitting-enriched QCD samples
  - selection:
    AK8 jet, p_T>425 GeV, **with soft muon**
    nsubjettiness, tau2/tau1<0.5: two-body decay



- Channel 2: **boosted top quarks**
  - semi-leptonic ttbar decays, muon channel
  - leptonic decay: isolated muon, p_T>50 GeV
  - hadronic decay: AK8 jet, p_T>400 GeV, tau3/tau2<0.86, softdrop mass [110, 210] GeV

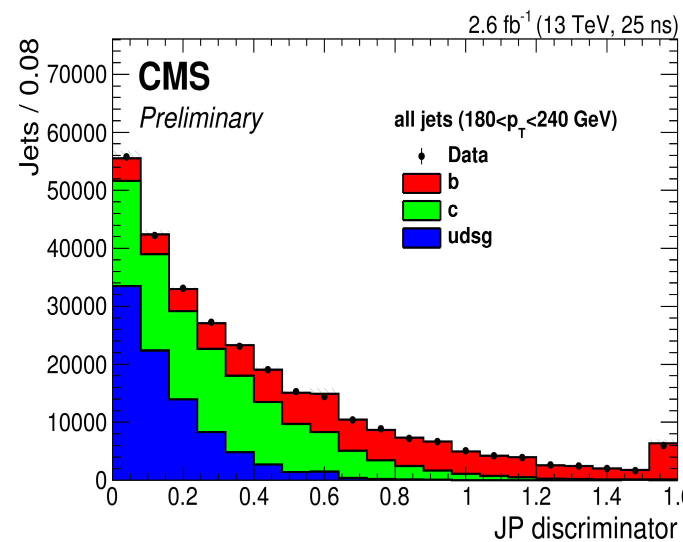


- Results are shown for subjet b tagging: same channels exploited to validate double b tagger, results released soon

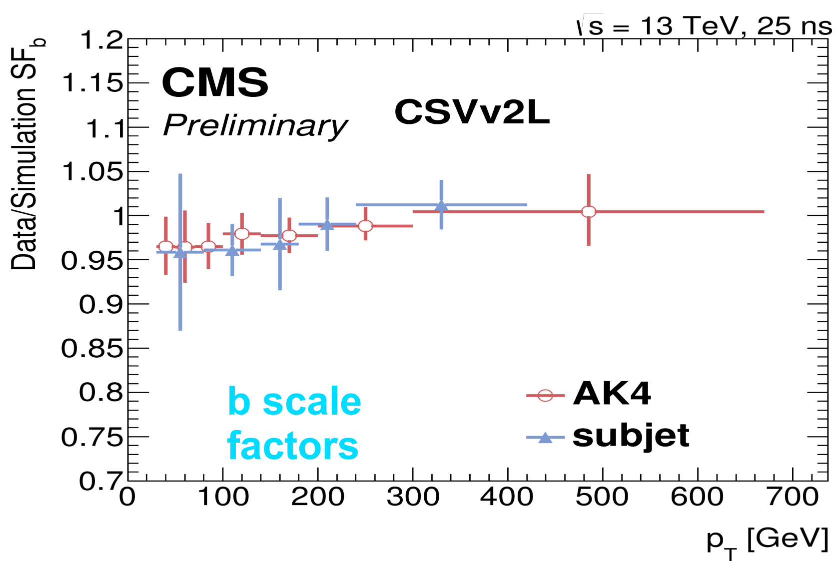
Boosted scale factors

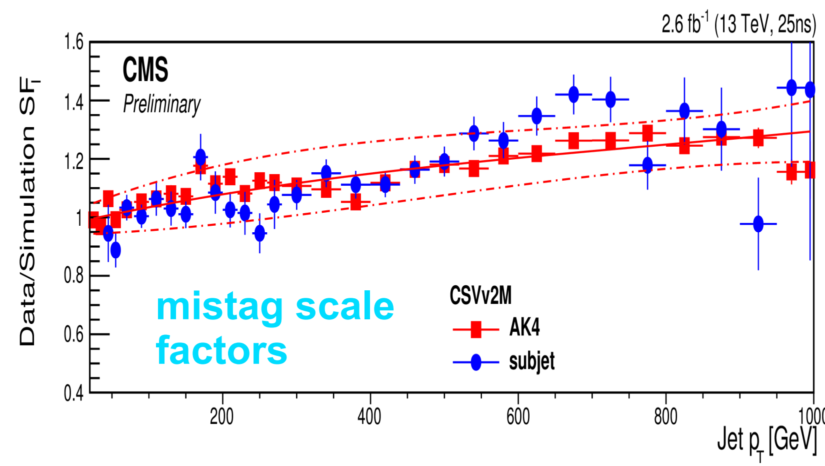
scale factors for **b subjets**

- Same life-time tagger (LT) method, as for AK4 jets
- **Template** fit concept based on **JP discriminator**:
  - JP has independent **calibration** in data and MC (mostly data-driven)
  - large fraction of b jets has JP information (>98%, for p_T>50 GeV)
- MC-based templates for b, c and light jets



scale factors for **light flavor subjets**: negative tag method





good agreement with non-boosted SF

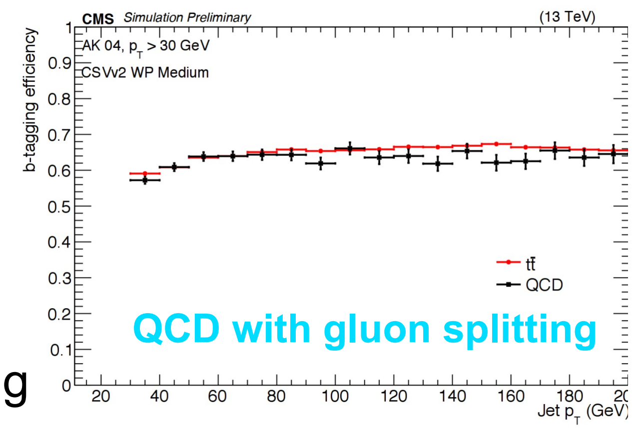
HF-LHC2016
items for discussion

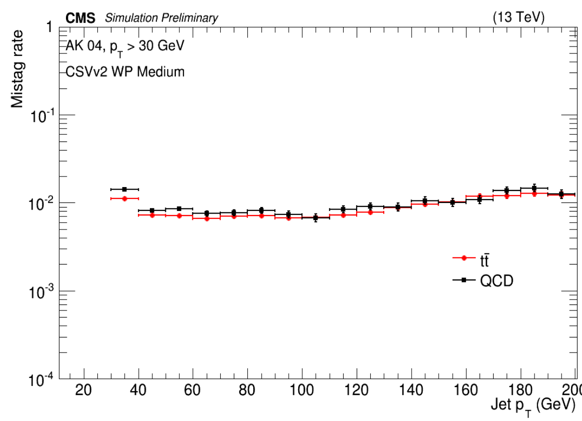
disclaimer: **many internal studies** ongoing, not shown here
detailed EvtGen investigation ongoing, but not made public yet

PDFs and gluon splitting

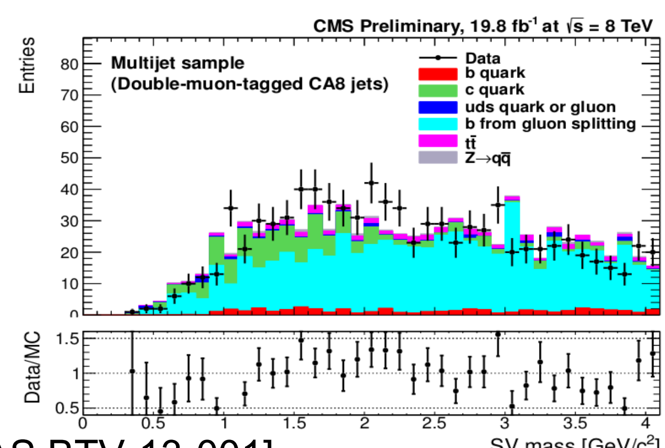
- Fit **PDFs and fragmentation functions simultaneously**?
  - having them **separated is probably ideal**, given the different sensitivity of the different analyses to the two sources of uncertainties
  - no clear user case known within the b tagging group

[CMS DP-2015/038]





- **Gluon splitting**:
  - **large impact** on the QCD performance
  - removing gluon splitting component **ttbar** and **QCD** performances diverge, still being fully understood:
    - **different content of the jet pT**: larger HF hadron contribution in ttbar events
    - **more gluons around b (c) jets** in QCD?
    - inputs welcome...
  - experience in getting gluon splitting enriched regions in data (e.g. boosted topologies studies)



[CMS PAS BTV-13-001]

Generator uncertainties

[https://twiki.cern.ch/twiki/bin/view/LHCPhysics/BTaggingSystematics]

- **Gluon splitting**:
  - fractions of jets with b-jets from gluon splitting varied by **+/- 50%**
  - impact:
    - low p_T: 0.1% - 0.3%
    - high p_T: 0.5% - 1.3%

- **Fragmentation function**:
  - p_T of the primary b-hadrons from b-quark fragmentation varied by **±5%**
  - impact: 0.2% - 0.8%

b/c prod.	low pT: 0.1% - 0.3%, high pT: 0.5% - 1.3%	QCD
mu pT	low pT: 0.1% - 1.1%, high pT: 0.1 - 0.9%	QCD
c/l ratio	<0.1% - 0.2%	QCD
b-frag	0.2% - 0.8%	QCD
PS (*)	0.3% - 0.6%	ttbar
IFSR	0.3% - 0.6%	ttbar

(* parton shower)

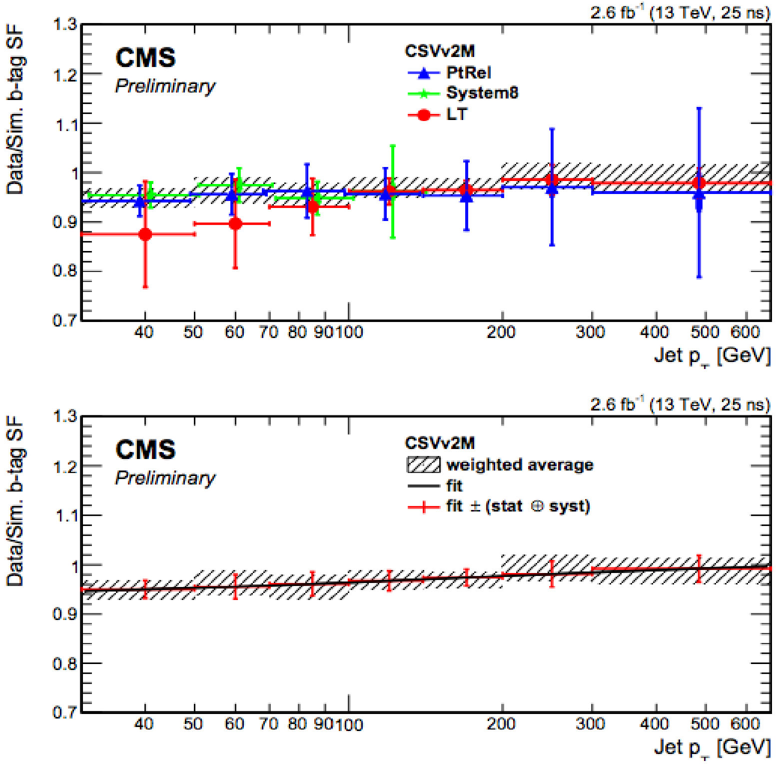
- **Branching ratios for D → μX, c → D fragmentation rate, $K_s^0(\Lambda)$ production fraction** are taken into account

- We do **not** calculate an uncertainty for **differences between generators**, because it would completely dominate the total uncertainty, but make sure that all generators are interfaced with **Pythia8** and measure SF with respect to that

- Further inputs on systematics treatment welcome

EvtGen

- We are investigating the use of EvtGen. Overall, it seems that **EvtGen does not improve data/MC** agreement: change in efficiency in opposite direction expected from p_T-dependent scale factors



[CMS PAS BTV-15-001]

- Importance of modeling of the **B hadron mass** and the **flight distance**:
  reweighting of flight distance significance (or mass) for b jets in Pythia8+EvtGen to that from Pythia8 removes most b tagging related discrepancies between generators

Other items

- How to use **PDG data**, in particular when the measurements are not necessarily consistent, e.g. if sum of exclusive modes > inclusive width?
  - the **exclusive modes** with a large branching fraction that are precisely measured, should be included as such
  - for the measurements with large branching fraction and large uncertainties and/or small branching fraction, but very different B/D-decay properties, implying a large impact, **variations** **could be proposed to obtain a systematic uncertainty**

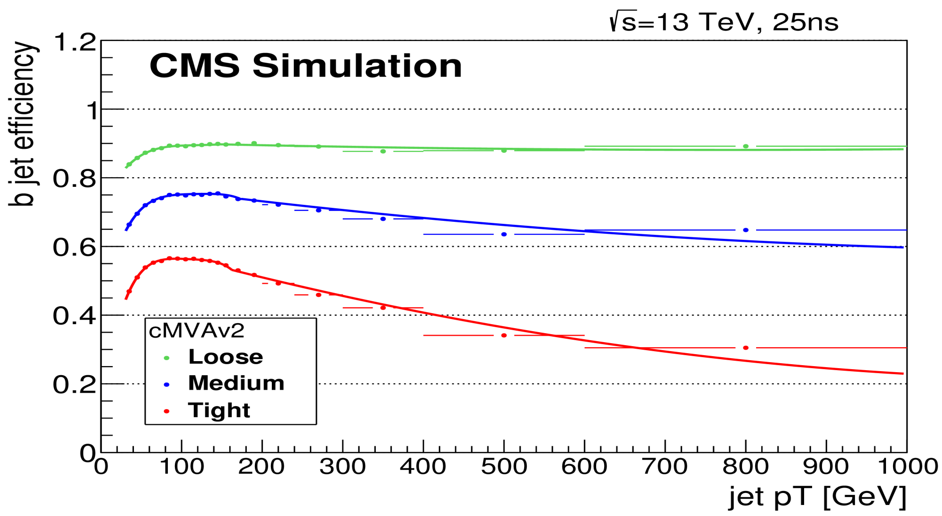
- LHCb has recently identified **several bugs in (EvtGen) b-decay models**. Have there been other experiment issues along these lines? How to validate generators?
  - we are not aware of additional **bugs** identified by CMS
  - **validation** can proceed exploiting both data and MC comparisons:
    compare events produced with different generators
    compare to measurements (whenever possible)

Outlook

- Overview of b tagging at CMS. Not covered important new results, some about to be released in public documents:
  - updated results on boosted topologies
  - **first CMS charm tagger**
  - **first CMS charge b/bbar tagger**

- Additional items on which we welcome feedback:
  - *difference between generators*
  - *inputs on our **RIVET** generation studies:*
    the idea is to define in RIVET a generator level selection close to the one used for b-tagging measurements, in order to perform the studies of different tunes and generator
  - *impact on b tagging due to the **massless** b quark assumption in some MC samples*

happy to provide **material for theory** studies

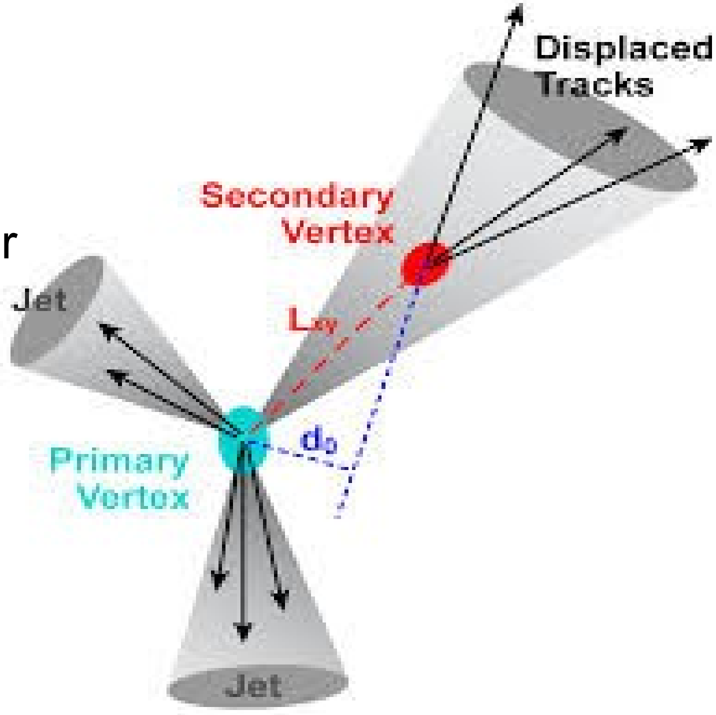


[CMS PAS BTV-15-001]

Additional Slides

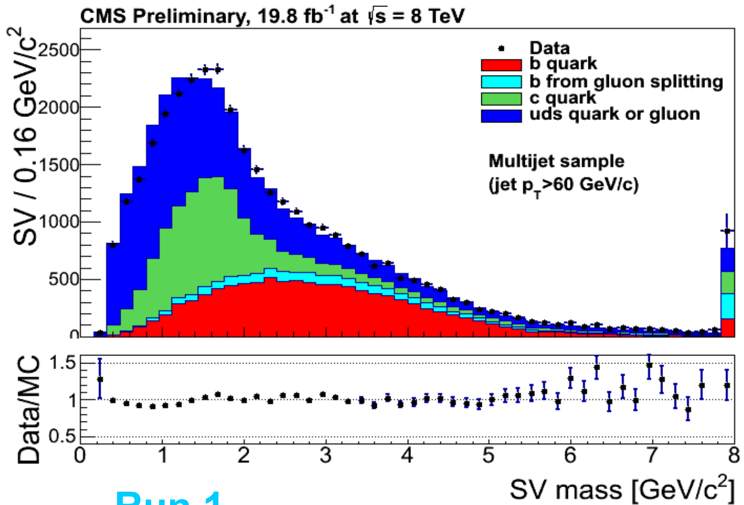
Introduction

- Standard b tagging at CMS: determine wether **AK4 jet** contains b hadron
- Exploits **properties** of b hadrons
  - **lifetime** ($c\tau$~500μm vs primary vertex resolution ~tens of μm)
  - **mass**: (~5 GeV)
  - decays with large **track multiplicities** (~5 tracks)
  - large **semileptonic** branching fraction (up to ~20% for both decays with electron or muon)
  - hard **fragmentation function**
- b tagging observables based on **track reconstruction**:
  - displaced tracks
  - secondary vertices
  - soft leptons
  - multivariate combination of the above
- algorithms produce **discriminator** values per each jet: large value → b-like jet

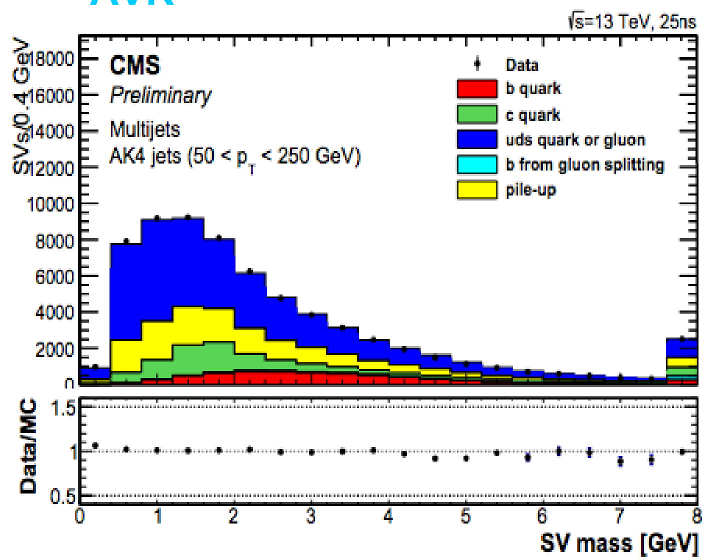


Secondary vertex reconstruction

- Adaptive vertex reconstruction (AVR) algorithm
  - default algorithm for Run 1 b tagging
  - starts from **tracks associated to the jets**
  - based on the adaptive vertex fitter
  - several selection criteria applied to remove secondary vertices less likely to originate from displaced B meson decays
- Inclusive vertex finder (IVF) algorithm
  - starts from **all tracks in the event**, no prior jet-track association
  - seeds for SV fit are displaced tracks with IP>50 μm and IP significance > 1.2
  - tracks in common with the event primary vertex are arbitrated, and the secondary vertex is refitted if at least two tracks remain
- Secondary vertex reconstruction in Run 2:
  - IVF is the default algorithm used for b tagging on AK4 jets and boosted topologies
  - **one multivariate algorithm (cMVAv2) exploits both IVF and AVR**



Run 1
AVR



Run 2 IVF

Working Points

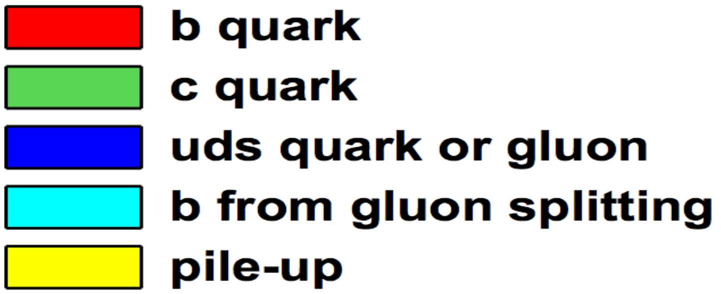
- Three working points are defined, to be used by physics analyses, defined as the cut on the discriminator value allowing to reduce the misidentification probability for light jets to definite values
  - **loose**, **medium** and **tight** working points correspond to misidentification probabilities of **10**, **1**, and **0.1 %**, respectively, based on QCD simulation
  - the evaluation of the efficiency is based on **ttbar events, jet p_T > 30 GeV**

Tagger	operating point	discriminator value	ϵ_b (%)
	JPL	0.245	≈ 82
JetProbability (JP)	JPM	0.515	≈ 62
	JPT	0.760	≈ 42
	CSVv2L	0.460	≈ 83
Combined Secondary Vertex (CSVv2)	CSVv2M	0.800	≈ 69
	CSVv2T	0.935	≈ 49
	cMVAv2L	-0.715	≈ 88
Combined MVA (cMVAv2)	cMVAv2M	0.185	≈ 72
	cMVAv2T	0.875	≈ 53

Track observables

- Standard b tagging track selection:
  - Transverse momentum p_T>1 GeV
  - Normalized χ^2<5
  - At least eight hits in the silicon tracker
  - At least two hits in the pixel layers of the tacker
  - Transverse impact parameter IP_{xy}<0.2 cm
  - Longitudinal impact parameter IP_z<17 cm
  - Distance to the jet axis D<0.07 cm
  - Decay length L<5 cm

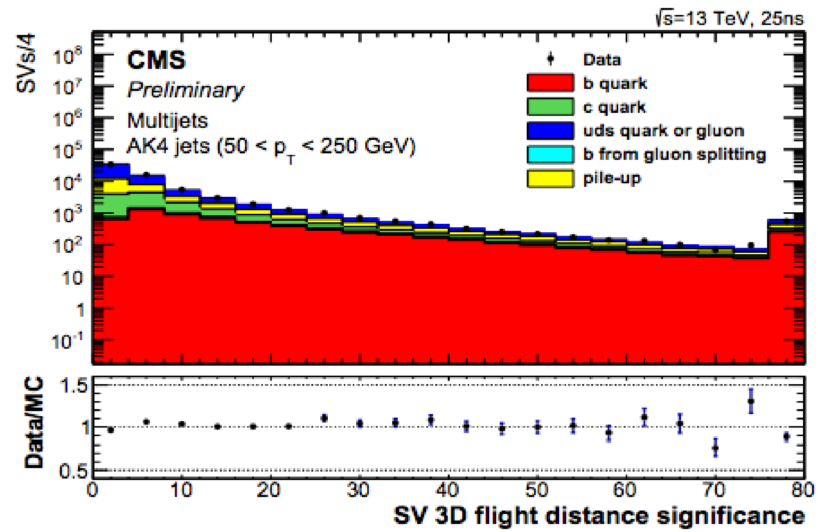
- MC contribution divided in flavors:

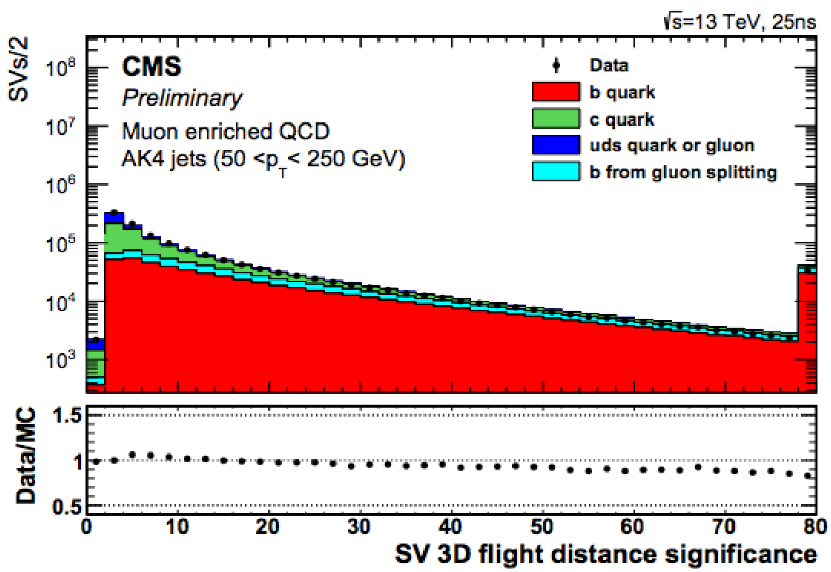


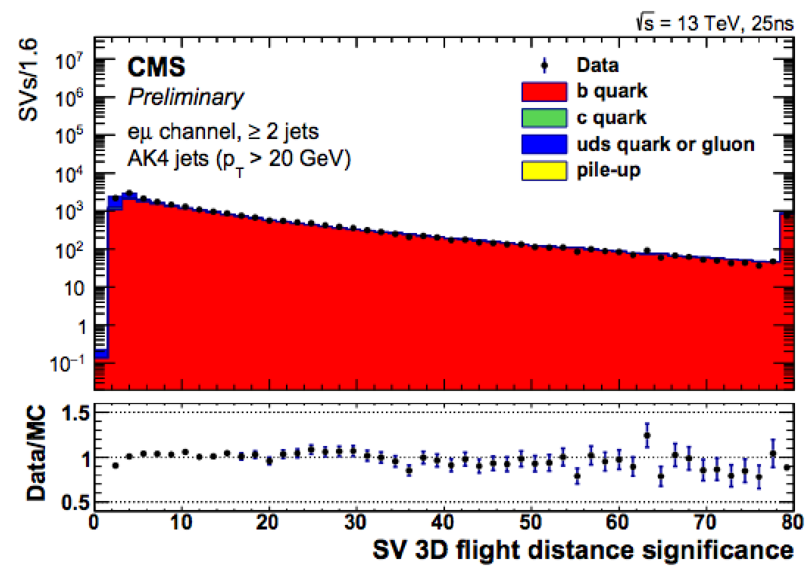
- For the muon-enriched QCD channel the pile-up contribution is negligible, thus absent in the legend

SV flight distance significance

SV flight distance significance for jets with an associated IVF secondary vertex , for the three event topologies

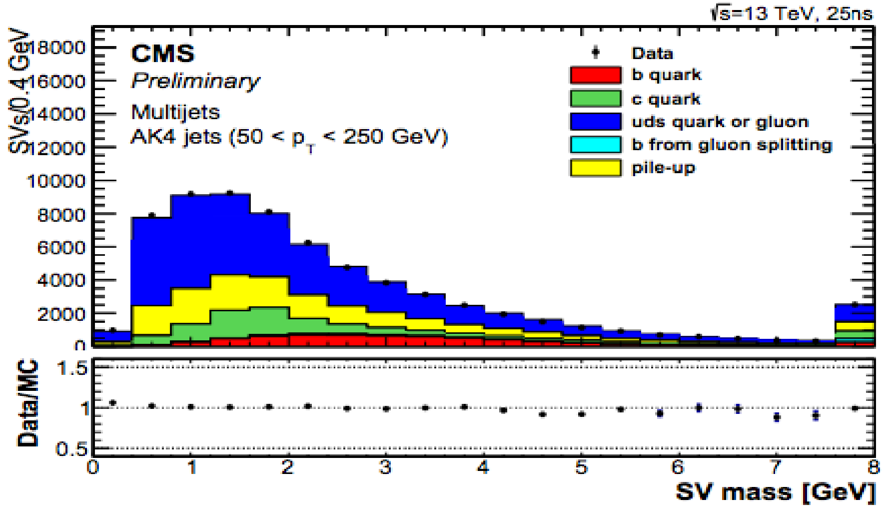


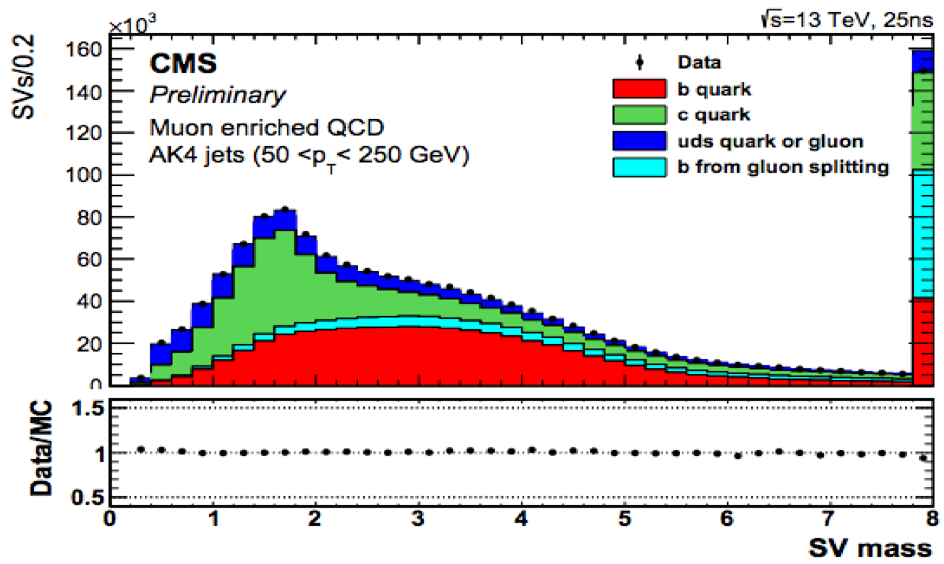


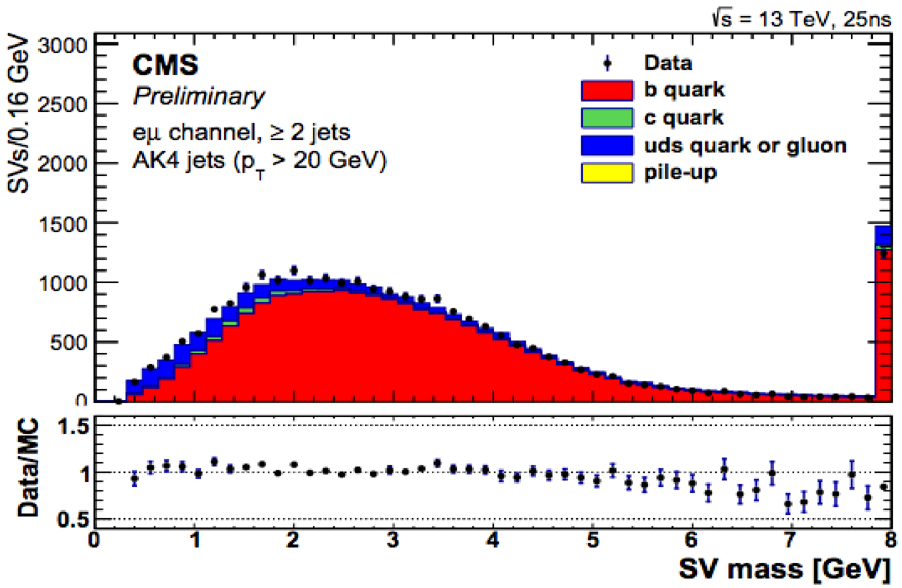


SV mass

SV mass for jets with an associated IVF secondary vertex, for the three event topologies







Combined MVA algorithm

- cMVAv2 algorithm
  - new algorithm developed in Run 2
  - it combines in a boosted decision tree (BDT) the discriminator values from six other algorithms:
    - **JP** and **JBP** taggers:
      the JBP tagger is a modified version of JP, using only the four tracks with highest impact parameter significance
    - **CSVv2(IVF)** and **CSVv2(AVR)**:
      as shown in the previous slides they are not fully correlated
    - Soft Muon (**SM**) and Soft Electron (**SE**) taggers:
      both algorithms are based on a BDT combination of discriminating variables such as the 2D and 3D impact parameter significances of the lepton-track and other lepton kinematic observables

Combination of scale factors: BLUE

- Results from PtRel, LT and System8 methods are combined using the a least squared **BLUE fit**:
  - same method used in Run1

- Proper treatment of **correlations** and anti-correlations of uncertainties between methods
  - statistical uncertainties are partially correlated according to the **fraction of data shared** by the different methods

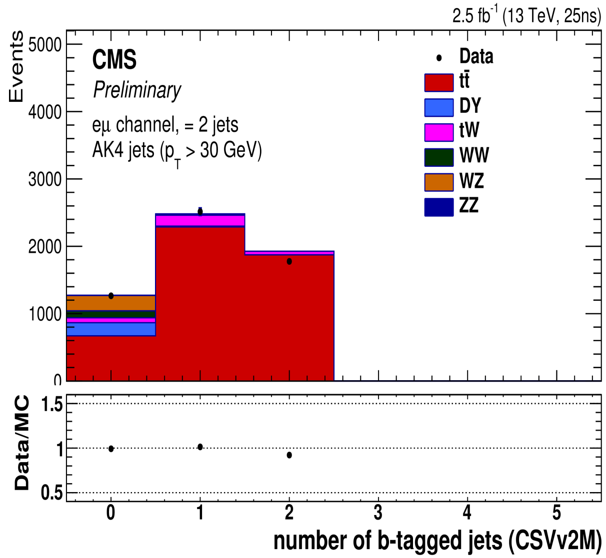
- A unique fit is done assessing **all the p_T bins** at the same time:
  - allows to correlate the systematic uncertainties across different bins
  - see: Nuclear Instruments and Methods in Physics Research A 500 (2003) 391

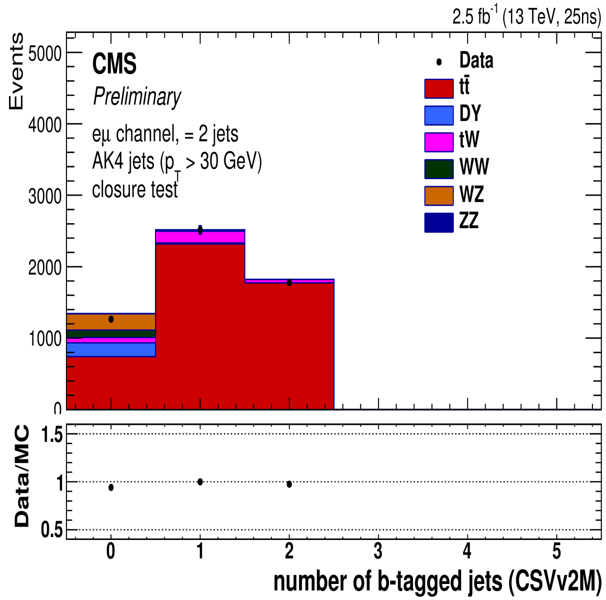
Tag Counting method in ttbar

- ttbar-enriched selection
  - ==2 jets (p_T> 30 GeV), **electron-muon** dilepton final state
- Simple but robust method: compare fraction of **events with 2 b tags** in data and MC:
  - subtract residual background
  - efficiency in data given by:

$$\varepsilon_b = \sqrt{\frac{F_{2btag} - F_{2btag}^{non-b-jet}}{f_{2b}}}$$

- Major sources of **systematics**:
  - 100% uncertainty assigned to the fraction of non-b-jets
  - 50% uncertainty assigned to background normalization





closure test: measured SF applied

Discriminator reweighting

- Designed for analyses exploiting the discriminator shape (e.g. in an MVA)

- Reweighting factors for **light- and b-jets** are simultaneously determined from iterative procedure on two samples

  - b-enriched sample, **ttbar dilepton**
  - ee, eμ, μμ channels
  - $|M_{ll} – M_{Z}| > 10$ GeV, $E_{T}^{miss}>30$ GeV
  - exactly two jets, $p_{T} > 20$ GeV
  - one tag jet, CSVv2M tagged (cMVAv2M tagged, when method applied to cMVAv2)

  - light-enriched sample, **Z→two leptons**
  - two same flavor leptons
  - $|M_{ll} – M_{Z}| < 10$ GeV, $E_{T}^{miss}<30$ GeV
  - exactly two jets, $p_{T} > 20$ GeV
  - one tag jet failing CSVv2L (cMVAv2L, when method applied to cMVAv2)

- Uncertainties include contamination of different flavors in each of the samples used, simulation statistics and jet energy scale

Generator uncertainties

[https://twiki.cern.ch/twiki/bin/view/LHCPhysics/BTaggingSystematics]

b/c production	b,c --> gg scale by 50%
B decay	neglected
b-quark frag.	av. B hadron energy fraction varied by +/- 5%
c/l ratio	l/c ratio scaled by 20%
muon pT	vary cut on muon pT
top generator	compare fit to templates for QCD
parton shower	comoare HERWIG to PYTHIA
ISR / FSR	varying Q2 scale and ME-PS threshold
underl. event	varing parameters